Population genetics of translational robustness

Claus O. Wilke* and D. Allan Drummond†

*Section of Integrative Biology
and Center for Computational Biology and Bioinformatics,
University of Texas, Austin, TX 78712, USA
†Program in Computation and Neural Systems,
California Institute of Technology,
Pasadena, CA 91125, USA

February 9, 2008

Running head: Translational robustness

Keywords: translation, evolutionary rate, expression level, neutrality, protein evolution

Corresponding author:
Claus O. Wilke
Integrative Biology
#1 University Station – C0930
University of Texas, Austin, TX 78712, USA
cwilke@mail.utexas.edu
Phone: (512) 471 6028
Fax: (512) 471 3878

Abstract: Recent work has shown that expression level is the main predictor of a gene's evolutionary rate, and that more highly expressed genes evolve slower. A possible explanation for this observation is selection for proteins which fold properly despite mistranslation, in short selection for translational robustness. Translational robustness leads to the somewhat paradoxical prediction that highly expressed genes are extremely tolerant to missense substitutions but nevertheless evolve very slowly. Here, we study a simple theoretical model of translational robustness that allows us to gain analytic insight into how this paradoxical behavior arises.

INTRODUCTION

The increasing availability of whole-genome sequences from many different species has revealed a surprising fact: Different genes within the same organism evolve at dramatically different rates. For example, the evolutionary rates of the fastest- and slowest-evolving genes in *Saccharomyces cerevisiae* are separated by three orders of magnitude (Drummond *et al.* 2005). Because the dominant force shaping genome-wide patterns of evolutionary rate is most likely stabilizing selection, the evolutionary rates of genes should correlate with quantities that measure how important or otherwise constrained a gene is (Kimura 1983; Ohta 1992). A wide array of such quantities have been proposed, shown to correlate with evolutionary rate, and subsequently disputed. Examples include a gene's dispensability or essentiality (Hurst and Smith 1999; Hirsh and Fraser 2001; Jordan *et al.* 2002; Pal *et al.* 2003; Zhang and He 2005; Wall *et al.* 2005), its number of interaction partners (Fraser *et al.* 2002; Bloom and Adami 2003; Jordan *et al.* 2003; Hahn *et al.* 2004; Agrafioti *et al.* 2005), its length (Marais and Duret 2001), or its centrality in the protein interaction network (Hahn and Kern 2005). However, it seems that most importantly, the expression level (Pal *et al.* 2001; Rocha and Danchin 2004), or perhaps more accurately the frequency of translation events (Drummond *et al.* 2005, 2006), influence evolutionary rate.

Drummond *et al.* (2005) have recently introduced a theory for why highly expressed genes evolve slowly. Translation is error prone, and inactivated or misfolded proteins resulting from mistranslation impose substantial costs on the cell (Bucciantini *et al.* 2002), costs which increase with expression level. One way in which the cost associated with a highly expressed gene is reduced is translational accuracy (Akashi 1994, 2001), whereby the gene is

encoded with optimal codons whose translation is less error-prone than the translation of other codons. Translational accuracy can explain why the rate of synonymous evolution dS is correlated with expression level, codon adaptation index, or protein abundance. However, it cannot explain why the rate of non-synonymous evolution dN shows an even stronger correlation with these quantities (Drummond *et al.* 2005). Selection for translational accuracy can reduce the translational error rate by a factor of 4–9 (Precup and Parker 1987), but even optimally coded genes that are highly expressed may produce a large amount of erroneous polypeptides. Therefore, Drummond *et al.* (2005) suggest that highly expressed genes should be under additional selection to be tolerant to translation errors, that is, the polypeptides produced from these genes should fold properly even if they were erroneously translated. Recent work on protein biochemistry has shown that proteins can differ widely in their tolerance to missense substitutions, and that properly chosen point mutations can dramatically increase the tolerance of a protein to additional substitutions (Bloom *et al.* 2005). Drummond *et al.*'s hypothesis, referred to as selection for translational robustness, predicts a constraint on the non-synonymous rate of evolution, whereas selection for translational accuracy predicts primarily a constraint on the synonymous rate of evolution. We must assume that both selective constraints will operate on genes that are frequently translated.

Genes that are highly tolerant to translational missense errors must also, by definition, be highly tolerant to missense substitutions. However, the translational robustness hypothesis predicts that these genes will nevertheless be strongly conserved under evolution. An example of a gene that exhibits this paradoxical behavior is Rubisco, an extremely abundant protein which fixes carbon dioxide during photosynthesis. Rubisco is strongly conserved

across phyla, but appears to tolerate most missense substitutions in the laboratory without loss of fold (Spreitzer 1993; Kellogg and Juliano 1997).

The purpose of the present paper is to put the translational robustness hypothesis into precise mathematical terms, and to demonstrate how highly expressed genes can evolve to be tolerant to missense substitutions and yet remain strongly conserved under evolution.

MATERIALS AND METHODS

Model: We consider the evolution of a gene encoding a protein of length L. Each site in the protein can be in one of two states, *neutral* or *non-neutral*. We denote the number of neutral sites in the protein by n. A mutation at a neutral site of a folded protein leaves the protein folded, but changes the site from neutral to non-neutral. A mutation at a non-neutral site of a folded protein will usually cause misfolding and consequent loss of function, but with a small probability α, such a mutation will not affect folding but turn the site into a neutral one. For simplicity, we assume that once an amino-acid sequence loses the ability to fold, it is impossible to mutate it back into a folded state. The rationale behind this assumption is that the likelihood of a mutation restoring fold to an unfolded amino-acid sequence is so low as to be negligible. Our model is a reasonable abstraction of a thermodynamic framework of protein evolution that has recently been shown to have good predictive power for both simulated and real proteins (Bloom *et al.* 2005; Wilke *et al.* 2005). The key insight of this framework is that a protein's tolerance to substitutions is closely related to the protein's stability—more stable proteins can withstand more missense substitutions—and that therefore proteins can change from being highly fragile to being highly robust to

mutations and vice versa through the accumulation of stabilizing or destabilizing mutations. In this sense, a mutation from a non-neutral to a neutral site in our model corresponds to a stabilizing mutation, and the opposite mutation corresponds to a destabilizing mutation. Thus, our model captures the following key aspects of protein biochemistry: (i) Homologous proteins can vary widely in their tolerance to mutations, and individual point mutations can increase or decrease this tolerance. (ii) Mutations that increase a protein's tolerance are much rarer than mutations that decrease its tolerance. (iii) Highly tolerant proteins are extremely rare, while moderately tolerant proteins are abundant. (iv) Non-functional mutant proteins are likely to be misfolded.

The gene is expressed at a level that leads to the synthesis of x polypeptides. For simplicity, we assume that the total number of polypeptides translated per gene is proportional to the gene's expression level, and that the constant of proportionality is 1. Thus, x is also the expression level, measured in mRNA molecules per cell. Under translation, each site is mistranslated with probability τ (we neglect premature termination of the translation process). The probability that a single mRNA molecule is mistranslated and leads to a misfolded protein is $1 - [1 - \tau(1 - \alpha)]^{L-n} \approx \tau(L - n)$, where the approximation holds for $\alpha, \tau \ll 1$. Let $f = \tau(L - n)$ be the fraction of synthesized proteins that misfold. We assume that the expression level is regulated such that the number of folded proteins per gene, the protein abundance A, is held constant, regardless of the translational error rate. Then, $A = x(1 - f)$. The total number of misfolded polypeptides per gene follows as $xf = Af/(1 - f)$. Finally, we assume that each misfolded polypeptide imposes a cost c on the cell, so that the total cost of a gene translated at abundance level A is $cAf/(1 - f)$. We turn this cost into a fitness value by

assuming that each misfolded protein has the same relative effect on fitness. Then we can write the overall fitness of a gene with n neutral sites as

$$w_n = \exp\Big(- cA\frac{\tau(L-n)}{1-\tau(L-n)}\Big)\,. \tag{1}$$

Without loss of generality, we use $c = 1$ henceforth.

Simulations: We implemented a stochastic simulation of N sequences reproducing in discrete, non-overlapping generations. We employed standard Wright-Fisher sampling, that is, the probability that a sequence in generation $t+1$ is the offspring of a sequence at generation t is directly proportional to the latter's fitness.

We measured the evolutionary rate along the line of descent from the most recent common ancestor (MRCA) of the final population backwards in time, as described by Wilke (2004). Briefly, we let the simulated population evolve until the birth-time of the population's MRCA, designated t_0, exceeded a fixed equilibration time $t_{\rm equil}$ plus a time window $t_{\rm meas}$, $t_0 > t_{\rm equil} + t_{\rm meas}$. All quantities were measured on the equilibrated population during this latter time window. For all results reported, we chose $t_{\rm equil} = t_{\rm meas} = 400000$, $U = 0.001$, $\tau = 0.001$, and $L = 100$. At all parameter settings, we carried out 100 replicas and averaged results over all replicas.

ANALYTICAL RESULTS

Solution based on Sella-Hirsh theory: We can calculate the steady-state solution of our model using the analogy between evolutionary biology and statistical physics recently demonstrated by Sella and Hirsh (2005). The theory of Sella and Hirsh (2005) is applicable whenever the product of population size and mutation rate is much smaller than one, $NU \ll 1$. In this regime, the population is essentially homogeneous at all times, and can

be represented at any given point in time by a single sequence. We say the population is in state i if the dominant sequence in the population is sequence i. The key insight of Sella and Hirsh (2005) is that the probability p_i to find the population in state i is proportional to a function $F(i)$ (also called a Boltzmann factor) that depends only on the fitness of sequence i, the population size, and details of the mutation process. Thus, it follows that

$$p_i = F(i) / \sum_j F(j) \,, \tag{2}$$

where the sum runs over all possible sequences j. Once we have the probabilities p_i, we can calculate all observable quantities of interest, such as expected fitness and expected evolutionary rate, using standard probability theory (see also below).

As the fitness of a sequence in our model depends only on the sequence's number of neutral sites n, it is useful to lump all sequences with the same n into a single class, and calculate the probability p_n that the population is in a state represented by any sequence with n neutral sites. Since there are $\binom{L}{n}$ such sequences, we introduce $F'(n) = \binom{L}{n} F(n)$, where $F(n)$ is the appropriate Boltzmann factor for a single sequence with n neutral sites, and then calculate p_n as $p_n = F'(n) / \sum_{k=0}^{L} F'(k)$. In our case, $F'(n)$ is given by

$$F'(n) = \binom{L}{n} \exp[(2N-2)\ln(w_n) + n \ln(\alpha)] \tag{3}$$

with w_n as defined in Eq. (1). The second term in the exponential takes into account the asymmetry in the mutation process, that is, mutations that increase n are a factor α less likely to occur than mutations that decrease n (Sella and Hirsh 2005, supplementary text).

With the formalism outlined in the previous paragraphs, we can calculate

the expected number of neutral sites in the steady state $\mathrm{E}[n]$ as

$$\mathrm{E}[n] = \sum_{n=0}^{L} nF'(n) / \sum_{k=0}^{L} F'(k)\,, \tag{4}$$

and the expected fitness as

$$\mathrm{E}[w] = \sum_{n=0}^{L} w_n F'(n) / \sum_{k=0}^{L} F'(k)\,. \tag{5}$$

Note that the expected values are not taken over the population (which is assumed to be homogeneous), but over a long time-window in the steady state. Next we can calculate the evolutionary rate K, that is, the expected number of amino-acid substitutions per unit time that accumulate along the line of descent in an equilibrated population. We find

$$K = \sum_{n=0}^{L} NU[\alpha\pi(n \to n+1) + \pi(n \to n-1)]F'(n) / \sum_{k=0}^{L} F'(k)\,, \tag{6}$$

where $\pi(i \to j)$ is the probability of fixation of sequence j in background i (Sella and Hirsh 2005),

$$\pi(i \to j) = \frac{1 - e^{-2\ln(w_j/w_i)}}{1 - e^{-2N\ln(w_j/w_i)}}\,. \tag{7}$$

[Note that $\pi(L \to L+1) := 0$ and $\pi(0 \to -1) := 0$.]

We can simplify these expressions in the special cases that A is either very large or very small. After inserting Eq. (1) into Eq. (3), we have

$$F'(n) = \binom{L}{n} \exp\Big[-2NA\frac{\tau(L-n)}{1-\tau(L-n)} + n\ln(\alpha)\Big]\,, \tag{8}$$

where we have also made the approximation $N - 1 \approx N$. We will continue to use this approximation throughout the rest of the paper. From Eq. (8), we can see that the behavior of the system changes drastically depending on

whether the product NA is large or small. However, since the population size N is the same for all genes in a species while each gene's corresponding protein abundance A can vary over many orders of magnitude, in the following we assume that N is fixed and consider the limits of very small and very large A.

For $A \to 0$, the first term in the exponential disappears, and $F'(n)$ becomes simply $\binom{L}{n}\alpha^n$. Thus, we find

$$\begin{aligned} \mathrm{E}[n] &= \sum_{n=0}^{L} n \binom{L}{n} \alpha^n / \sum_{n=0}^{L} \binom{L}{n} \alpha^n \\ &= \alpha L/(\alpha+1)\,. \end{aligned} \tag{9}$$

We cannot obtain similarly simple expressions for $\mathrm{E}[w]$ and K in this limit, but will do so in the next subsection using a different method.

For $A \to \infty$, we have to distinguish between the cases $n = L$ and $n < L$. For $n < L$, the first term in the exponential in Eq. (8) becomes much larger in magnitude than the second term, which is a constant. Thus, we can neglect the second term, and have

$$F'(n) = \binom{L}{n} \exp\Big[-2NA \frac{\tau(L-n)}{1-\tau(L-n)} \Big]\,. \tag{10}$$

This expression tends to 0 for large A. For $n = L$, we have $F'(L) = \alpha^L$, independent of A. All terms but the $n = L$ term disappear, and we have $\mathrm{E}[n] = L$, $\mathrm{E}[w] = w_L$, and

$$\begin{aligned} K &= NU\pi(L \to L-1) \\ &= NUe^{-2NA\tau/(1-\tau)}\,. \end{aligned} \tag{11}$$

Approximate solution: Sella-Hirsh theory yields an exact solution for our model. However, the resulting expressions are somewhat unwieldy, and don't

lead to simple analytical expressions for intermediate A. Therefore, we will now derive an approximate solution to our model.

For small τ, we can approximate $w_i \approx \exp[-A\tau(L-i)]$, and find $\ln(w_j/w_i) = A\tau(j-i)$. This approximation is equivalent to neglecting the small number of additional translation events required to replace polypeptides that misfold. The probability of fixation follows from Eq. (7) as

$$\pi(i \to j) = \frac{1 - e^{-2A\tau(j-i)}}{1 - e^{-2NA\tau(j-i)}} \,. \tag{12}$$

The idea of the approximate solution is that in the steady state, mutations that increase the number of neutral sites and those that decrease it are in perfect balance. Therefore, the number of neutral sites in the steady state, n^*, satisfies:

$$\alpha(L - n^*)\pi(n^* \to n^* + 1) = n^*\pi(n^* \to n^* - 1) \,. \tag{13}$$

According to Eq. (12), $\pi(n \to n+1)$ and $\pi(n \to n-1)$ are independent of n. We introduce π_+ and π_-, the probabilities of fixation for a mutation that increases or decreases n by one, respectively, and find:

$$\pi_+ = \frac{1 - e^{-2A\tau}}{1 - e^{-2NA\tau}} \,, \tag{14}$$

$$\pi_- = \frac{1 - e^{2A\tau}}{1 - e^{2NA\tau}} \,. \tag{15}$$

After inserting these expressions into Eq. 13 and solving for n^*, we obtain

$$n^* = \frac{\alpha L \pi_+}{\alpha\pi_+ + \pi_-} \,. \tag{16}$$

With this result, the expected fitness in the steady state becomes

$$\mathrm{E}[w] = \exp[-A\tau(L - n^*)] \,, \tag{17}$$

and the evolutionary rate K follows as

$$K = NU\Big(\alpha\pi_+\frac{L-n^*}{L} + \pi_-\frac{n^*}{L}\Big)\,. \tag{18}$$

In the Appendix, we derive an expression for K as a function of n^*, rather than as a function of A as we have done here.

In the limit $A \to 0$, we have $\pi_+ = \pi_- = 1/N$ and $n^* = \alpha L/(\alpha+1)$. [Note that this expression is identical to the result found through Sella-Hirsh theory, Eq. (9), if we equate n^* with E[n].] Therefore, in this limit,

$$K = \alpha U\,. \tag{19}$$

In the limit $A \to \infty$, we have $\pi_+ = 1$, $\pi_- = e^{-2NA\tau}$, and $n^* = L$. Therefore, in this limit,

$$K = NUe^{-2NA\tau}\,. \tag{20}$$

The expressions for n^* and K again agree with the results found through Sella-Hirsh theory, if we assume $1-\tau \approx 1$ in Eq. (11).

Limitations on the number of neutral sites: Certain residues may never tolerate any substitutions, such as the active-site serine of a serine protease or the heme-binding histidines of hemoglobin. Under the assumption that m sites can never be neutral, we can write $L = l + m$, and for small τ the fitness w_n is approximately

$$w_n = e^{-A\tau(l+m-n)} = e^{-A\tau m}e^{-A\tau(l-n)}. \tag{21}$$

In other words, all fitness values are rescaled by a factor depending on τ but not on n. Eq. (7) reveals that such a rescaling leaves the fixation probabilities unchanged. Therefore, the approximate solution remains unchanged except that we replace L by l $(= L - m)$ everywhere and add a factor $e^{-A\tau m}$ to Eq. (17).

For Sella-Hirsh theory, if we assume that τm is negligibly small, the Boltzmann factor $F'(n)$ gains a similar leading factor which, as Eqs. (4) and (5) make clear, also drops out, this time through the normalization term. In the case of $\mathrm{E}[n]$, the result is only that L must again be replaced by $l = L - m$, while the expected value $\mathrm{E}[w]$ also gains a leading factor $e^{-A\tau m}$, just as in the approximate case.

In short, when there are m never-neutral sites, the main effects are to lower the population's fitness by a factor $e^{-A\tau m}$ and to reduce the expected number of neutral sites $\mathrm{E}[n]$ and the evolutionary rate K roughly as though the evolving gene were shorter by m residues.

SIMULATION RESULTS

First, we studied the rate of evolution K as a function of the protein abundance A, for various population sizes (Fig. 1). We found that K levels off for $A \to 0$. The asymptotic value at $A = 0$ is $K(0) = \alpha U$ (Eq. 19). For increasing A, K first increases, and then rapidly drops off to zero for even larger A. The main effect of the population size N is to determine at what level of A this drop-off initiates. With increasing N, the evolutionary rate K seems to be simply shifted to the left, towards lower A (Fig. 1). We can make this statement more precise by considering the large-A limit of our approximate solution, Eq. 20. This limit shows that the evolutionary rates $K(N, A)$ and $K(aN, a^{-1}A)$ are related through $K(N, A) \approx a^{-1}K(aN, a^{-1}A)$, where a is an arbitrary constant. Therefore, if we increase N by a factor of a, the resulting curve $K(A)$ appears on a log-log plot to be shifted upwards and to the left by an amount of $\log(a)$. The upwards shift cannot be noticed, because it exists only for very large A, and thus the effect of increasing N seems to be to simply shift the $K(A)$ curve to the left.

Second, we studied the effect of varying α on $K(A)$ (Fig. 2). The variable α mainly influences the asymptotic limit of $K(A)$ for small A (with $K(A)$ increasing with α), but does not affect how quickly $K(A)$ decays for large A.

Third, we studied the behavior of the expected fitness and the expected number of neutral sites for varying levels of A. The expected fitness is approximately 1 for both very low and very high A, but drops below 1 for the intermediate values of A for which $K(A)$ starts to decay (Fig. 3a). The expected number of neutral sites is $\alpha L/(\alpha+1)$ for very small A, and increases to 1 for large A (Fig. 3b). We can understand the different evolutionary regimes of low A and high A as follows: For low A, since very little protein is synthesized, the cost associated with misfolded proteins following erroneous translation is negligible. Therefore, the expected fitness in this regime is 1. Since the cost of misfolding is negligible, the number of neutral sites is not under selection in this regime, and it settles to the value at which the mutations increasing n and those decreasing n exactly balance each other. For high A, on the other hand, the cost of translation-induced misfolding is tremendous. Therefore, at high A, the population converges to the single optimal sequence with $n = L$ (or $n = l$ if some sites can never be neutral). Every mutation that reduces n by even 1 is highly deleterious, and therefore will virtually never go to fixation, even in a small population. For $l = L$, the optimal sequence (which has $n = l$) pays no cost whatsoever under mistranslation, and the expected fitness is again 1. For intermediate A, the cost of translation-induced misfolding is significant but not debilitating. As a result, n is elevated in comparison to its low-A limit, but the expected fitness falls below 1.

Finally, the calculation in the Appendix predicts that the evolutionary rate should be independent of N if we plot it as a function of the expected

number of neutral sites E[n] rather than a function of A. Figure 4 shows that this prediction is indeed accurate. We find that there are two distinct regimes for the evolutionary rate. For small E[n], the evolutionary rate increases with E[n]. This increase is caused by the increased availability of neutral mutations with growing E[n]. However, even though we can calculate what the evolutionary rate would be for arbitrarily small E[n], in equilibrium E[n] will never be below its limiting value for small A, $\alpha L/(\alpha+1)$. For large E[n], the behavior of K is reversed, and now it decreases with increasing E[n]. The decay comes about because in this regime, even though there are many mutations which do not disrupt the fold of a properly translated protein, these mutations increase the amount of mistranslated, misfolded proteins, and thus are selected against. The quantity E[n] can get arbitrarily close to L (or l for $m > 0$), and therefore K can decay to almost zero if A is sufficiently large.

Throughout this study, we found good agreement among the numerical simulations, Sella-Hirsh theory, and our simple analytical approximation. Some discrepancies appeared between theory and simulations for the largest population size ($N = 1000$) and for very small α in conjunction with large A. We attribute the former to a violation of the condition $NU \ll 1$, which must be satisfied for both Sella-Hirsh theory and our approximate solution. We carried out our simulations with a mutation rate of $U = 0.001$, which means that $NU = 1$ for $N = 1000$. The latter discrepancies are caused by insufficient equilibration time. For large A, the number of neutral sites n always approaches L, irrespective of population size or α. However, the smaller α is, the lower the probability that a mutation occurs which increases n. Therefore, the equlibration time needed at large A grows without bound as α decreases. We did additional simulations in this regime, and found that

the simulation results approached the predicted quantities with increasing equilibration time (data not shown).

DISCUSSION

We have developed a simple model that describes the slowdown of the rate of evolution of highly expressed genes under selective pressure for translational robustness. We have studied the model with numerical simulations and have solved the model exactly using Sella-Hirsh theory. We have also developed a simple analytic approximation that is in excellent agreement with the predictions from Sella-Hirsh theory, and is valid for the entire range of possible parameter values (as long as $NU \ll 1$).

The model abstracts a previous thermodynamic model of protein mutational tolerance introduced by Bloom *et al.* (2005) in which mutations may leave unperturbed or destabilize the protein's native structure (common) or stabilize it (uncommon). Increases in stability tend to increase the number of sites at which substitutions can be tolerated, so-called neutral sites, while decreases in stability usually cause misfolding or decrease the number of neutral sites. In the present work, we have modeled neutral sites directly. In doing so, we only allow stepwise changes in neutral sites, sacrificing treatment of large stability changes that might radically alter the number of neutral sites and the potential stability dependence of mutational effects in order to gain analytical tractability.

Our results show a clear example of the paradox cited in the Introduction (Fig. 4): Given selection against the costs of protein misfolding, genes simultaneously become more mutationally tolerant (larger number of neutral sites, n) but evolve slower as if fewer mutations were tolerated. The paradox's resolution requires disentangling two kinds of mutational tolerance, one

of which captures the likelihood of loss of protein function due to a mutation while the other quantifies the fitness cost of that mutation, the cost which ultimately determines evolutionary rate. When protein misfolding imposes minimal fitness costs, as is the case with low-expression proteins, the proportion of mutations which preserve protein function govern the rate of evolution (Fig. 4, left). However, at high expression levels, fitness costs of mutations which preserve protein function grow large and can become the dominant determinant of evolutionary rate (Fig. 4, right).

This observation has an important corollary. When selection for translational robustness is weak, functional loss is likely the main determinant of fitness costs. Thus, our results suggest that evolutionary conservation of sites in low-expression proteins may be more likely to indicate functional importance than similar conservation at sites in high-expression proteins.

Our simple model produces an exponential decline in evolutionary rate with increasing expression level, whereas in yeast, a power law better describes this relationship (Drummond *et al.* 2005). Several possibilities may explain the discrepancy. First, our model assumes a symmetric binomial distribution of the number of neutral sites, but the distribution for real proteins may be skewed or heavy-tailed. Second, the cost of additional misfolded proteins may not be independent of the number of already misfolded proteins. For example, misfolded proteins form toxic aggregates (Bucciantini *et al.* 2002), and aggregation is not a linear function of protein concentration. Finally, differences in protein structure and function between high- and low-expression proteins may play a role. Drummond *et al.* (2005) have previously examined differences between functionally and structurally similar paralogs and found a similar power-law relationship. However, more subtle but important differences may separate paralogs and influence their evolu-

tion.

Our results here demonstrate that profound differences in protein evolutionary rates can arise even in the absence of functional and structural differences and when variables such as protein length, the translation error rate, and the underlying distribution of the number of neutral sites are held constant. In real genomes, of course, all these features vary and some, perhaps all, are under selection. The value of the model is its utility in explaining why highly expressed proteins evolve slowly across taxa (Drummond *et al.* 2005).

Interestingly, our model reveals two evolutionary-rate regimes (Figs. 1 and 2), one in which rates remain relatively constant with increasing protein production, and another in which rates decline precipitously. In yeast, virtually all genes appear to fall in the latter regime, as the evolutionary rate declines almost immediately from low to high expression (Drummond *et al.* 2005), raising the possibility of genome-level selection in this direction. If yeast protein synthesis levels reflect organismal needs and cannot be freely modulated, as seems likely, and protein synthesis costs dominate cellular energy consumption, as evidence suggests (Princiotta *et al.* 2003), the remaining genome-level target for selection is on the fitness cost per misfolded protein, c. Decreasing c pushes genes away from the decline to where cost differences become negligible, whereas increasing c pushes genes toward the decline, amplifying the cost difference between high- and low-expression proteins. One way to decrease c is to drive translation errors down to negligible levels. Another is to maintain a quality-control apparatus (e.g. chaperones and proteases) with so much excess bandwidth that cost differences associated with variability in protein misfolding become negligible. Evidence suggests that both strategies for decreasing c impose significant fitness penalties.

Decreasing translational error rates can be easily accomplished, often with a single ribosomal mutation (Alksne *et al.* 1993), but ribosomal accuracy and growth rate are often negatively correlated, presumably through the intrinsic speed/accuracy tradeoff inherent in ribosomal proofreading (Kurland 1992). Maintenance of a chaperone fleet large enough to dilute out misfolding cost differences would divert enormous cellular resources for little benefit, and the massive induction of chaperones after heat shock suggests that cellular chaperone levels do not have much remaining bandwidth under normal conditions. Overall, it seems plausible that the steep decline observed in yeast's evolutionary-rate–expression relationship reflects a balance favoring a relatively high cost per misfolded protein c. Costly translational accuracy and quality control machinery may be reduced so long as the increased errors and reduced folding assistance can be compensated for; translational robustness provides that compensation—essentially for free—but is ultimately limited by mutation pressure away from robust sequences and by the fundamental intolerance of proteins to at least some errors.

Our model distinguishes between the number of polypeptides produced per gene, x, and the abundance of functional proteins, A, yet our approximate solution essentially equates these quantities with only minor accuracy loss. The approximation works for two reasons. First, misfolded polypeptides impose a negligible cost for low-abundance proteins, while for high-abundance proteins, misfolded polypeptides are rare because of selection for translational robustness. We expect these nontrivial results to hold for many organisms. Second, in our model, the number of translation events x, the primary determinant of the number of mistranslated proteins, is estimated accurately by A, a situation unlikely to hold for most organisms. Protein abundance reflects a balance of ongoing translation and turnover

(Greenbaum *et al.* 2003), such that a high abundance can result from either moderate translational frequency and long protein half-life or from rapid translation and short half-life. Because half-lives can vary over orders of magnitude (Hargrove and Schmidt 1989), abundance and translation frequency may only weakly correlate in real organisms. Among protein abundance, mRNA expression level, and translation frequency, we hypothesize that the latter, even though difficult to measure, will best predict evolutionary rate.

Bürger *et al.* (2005) recently studied a question closely related to the present paper, asking why phenotypic mutation rates (corresponding to the translational error rate in the present paper) are much higher than genotypic mutation rates. Within the framework of their model, Bürger *et al.* (2005) found very little pressure for reduction of phenotypic mutation rates below a certain threshold. Even though we keep the translational error rate constant in our model, we can consider a change in the number of neutral sites n as a change in the phenotypic mutation rate, and thus compare our results to those of Bürger *et al.* (2005). In contrast to their conclusions, we find that the fraction of neutral sites, n/L, quickly rises to the maximum possible for highly expressed genes, thus reducing the phenotypic mutation rate to zero except when some sites cannot be made neutral. We believe that the differences in results are caused by differences in the way in which we and Bürger *et al.* (2005) treated costs of erroneously translated proteins in our models. Bürger *et al.* (2005) consider the total cost of protein synthesis, but do not consider additional penalties imposed by misfolded proteins, not only for their recognition and cleanup by the quality-control system but also for their innate toxicity (Bucciantini *et al.* 2002). Clearly, if we neglect these unique costs, then the only pressure to reduce the phenotypic mutation

rate is to reduce the cost of synthesis for misfolded proteins, and this pressure will be weak if this cost is only a small proportion of the total cost of protein synthesis. In our model, on the other hand, we have focused exclusively on costs of misfolded proteins apart from their synthesis costs, implicitly assuming that the total cost of protein synthesis is approximately equal to the cost of synthesis of functional proteins, and that the benefit of the functional proteins will pay for their synthesis. We believe that there is indeed a strong selective pressure to reduce the phenotypic mutation rate for highly expressed genes, but that it is cheaper for cells to evolve translationally robust genes than to evolve highly accurate transcription and translation machinery.

Can translational robustness really be obtained cheaply? DRUMMOND *et al.* (2005) have suggested that increased protein stability both confers mutational tolerance and constrains sequence evolution. Increasing protein stability, that is, decreasing the free energy of folding ΔG_f, provides a plausible mechanism for obtaining translational robustness for numerous reasons. First, increased stability leads to increased mutational tolerance and can be obtained by point mutations (BLOOM *et al.* 2005). Second, the stability-increase mechanism is sufficiently general to encompass proteins of diverse functions and to operate in a wide range of organisms. Third, stability is free in the sense that obtaining a protein with lower ΔG_f requires only a chance mutation. While many researchers have noted an apparent tradeoff between protein stability and enzymatic activity, it is crucial to emphasize that this trend may be statistical rather than intrinsic: Because both high activity and high stability are rare properties, mutations that improve both are exceedingly unlikely unless both are constrained (GIVER *et al.* 1998). Selection for translational robustness provides precisely that dual constraint, and because many millions of mutations may be screened over evolutionary time,

the very few resulting in highly expressed proteins with increased stability (conferring tolerance to translation errors) and uncompromised activity will be found. Finally, the very rarity of such stabilizing mutations provides a measure of the scarcity of highly stable proteins available for exploration by evolutionary drift. If increased stability is a dominant response to the need for mutational tolerance in highly expressed proteins, it will restrict drift and slow evolution relative to less-constrained low-expression proteins.

ACKNOWLEDGMENTS

C.O.W. was supported by NIH grant AI 065960 and D.A.D. was supported by NIH National Research Service Award 5 T32 MH19138. D.A.D. acknowledges, with gratitude, the support of Frances Arnold.

References

Agrafioti, I., J. Swire, J. Abbott, D. Huntley, S. Butcher and M. P. H. Stumpf, 2005 Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. BMC Evol. Biol. **5:** 23.

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics **136:** 927-935.

Akashi, H., 2001 Gene expression and molecular evolution. Current Opinion in Genetics & Development **11:** 660-666.

Alksne, L. E., R. A. Anthony, S. W. Liebman and J. R. Warner, 1993 An accuracy center in the ribosome conserved over 2 billion years. Proc. Natl. Acad. Sci. USA **90:** 9538-9541.

Bloom, J. D. and C. Adami, 2003 Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol. Biol. **3:** 21.

Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami and F. H. Arnold, 2005 Thermodynamic prediction of protein neutrality. Proc. Natl. Acad. Sci. USA **102:** 606-611.

Bucciantini, M., E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C. M. Dobson and M. Stefani, 2002 Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature **416:** 507-511.

Bürger, R., M. Willensdorfer and M. A. Nowak, 2005 Why are phenotypic mutation rates much higher than genotypic mutation rates? Genetics, in press. doi:10.1534/genetics.105.046599.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke and F. H. Arnold, 2005 Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA **102:** 14338-14343.

Drummond, D. A., A. Raval and C. O. Wilke, 2006 A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol., in press.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman, 2002 Evolutionary rate in the protein interaction network. Science **296:** 750-752.

Giver, L., A. Gershenson, P.-O. Freskgard and F. H. Arnold, 1998

Directed evolution of a thermostable esterase. Proc. Natl. Acad. Sci. USA **95:** 12809-12813.

Greenbaum, D., C. Colangelo, K. Williams and M. Gerstein, 2003 Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. **4:** 117.

Hahn, M. W., G. C. Conant and A. Wagner, 2004 Molecular evolution in large genetic networks: Does connectivity equal constraint? J. Mol. Evol. **58:** 203-211.

Hahn, M. W. and A. D. Kern, 2005 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. **22:** 803-806.

Hargrove, J. L. and F. H. Schmidt, 1989 The role of mRNA and protein stability in gene expression. FASEB J. **3:** 2360-2370.

Hirsh, A. E. and H. B. Fraser, 2001 Protein dispensability and rate of evolution. Nature **411:** 1046-1049.

Hurst, L. D. and N. G. C. Smith, 1999 Do essential genes evolve slowly? Curr. Biol. **9:** 747-750.

Jordan, I. K., I. B. Rogozin, Y. I. Wolf and E. V. Koonin, 2002 Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12:** 962-968.

Jordan, I. K., Y. I. Wolf and E. V. Koonin, 2003 No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol. Biol. **3:** 1.

Kellogg, E. and N. Juliano, 1997 The structure and function of RuBisCO and their implications for systematic studies. Am. J. Botany **84:** 413-428.

Kimura, M., 1983 The neutral theory of molecular evolution. Cambridge University Press.

Kurland, C. G., 1992 Translational accuracy and the fitness of bacteria. Annu Rev Genet **26:** 29-50.

Marais, G. and L. Duret, 2001 Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. **52:** 275-280.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263-286.

Pal, C., B. Papp and L. D. Hurst, 2001 Highly expressed genes in yeast evolve slowly. Genetics **158:** 927-931.

Pal, C., B. Papp and L. D. Hurst, 2003 Rate of evolution and gene dispensability. Nature **421:** 496-497.

Precup, J. and J. Parker, 1987 Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. **262:** 11351-11355.

Princiotta, M. F., D. Finzi, S. B. Qian, J. Gibbs, S. Schuchmann, F. Buttgereit, J. R. Bennink and J. W. Yewdell, 2003 Quantitating protein synthesis, degradation, and endogenous antigen processing. Immunity **18:** 343-354.

Rocha, E. P. C. and A. Danchin, 2004 An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. **21:** 108-116.

Sella, G. and A. E. Hirsh, 2005 The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA **102:** 9541-9546.

Spreitzer, R. J., 1993 Genetic dissection of Rubisco structure and function. Annu. Rev. Plant Physiol. Plant Mol. Biol. **44:** 411-434.

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen and M. W. Feldman, 2005 Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA **102:** 5483-5488.

Wilke, C. O., 2004 Molecular clock in neutral protein evolution. BMC Genetics **5:** 25.

Wilke, C. O., J. D. Bloom, D. A. Drummond and A. Raval, 2005 Predicting the tolerance of proteins to random amino acid substitution. Biophys. J. **89:** 3714-3720.

Zhang, J. and X. He, 2005 Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol. Biol. Evol. **22:** 1147-1155.

APPENDIX

Here we derive an expression for the evolutionary rate K as a function of the number of neutral sites $n*$ rather than the protein abundance A. For the remainder of this appendix, we drop the superscript from n^* for simplicity. We begin by noting that Eq. (16) implies

$$n\pi_- = \alpha(L-n)\pi_+ \,. \tag{22}$$

Further, note that we can write, for sufficiently large N,

$$\pi_- = e^{-2NA\tau}\pi_+ \,. \tag{23}$$

Therefore, $e^{-2NA\tau} = \alpha(L-n)/n$. We can solve this expression for A and find

$$A = \frac{-1}{2N\tau} \ln[\alpha(L-n)/n] \,. \tag{24}$$

After inserting Eq. (24) into the definition of π_-, we obtain

$$\pi_- = \frac{1 - \exp\left(\frac{-1}{N} \ln[\alpha(L-n)/n]\right)}{1 - \exp\left(-\ln[\alpha(L-n)/n]\right)} \,. \tag{25}$$

We expand this expression for large N, replacing the exponential in the numerator by the first two terms of its Taylor series, and find

$$\pi_- = \frac{1}{N} \frac{\alpha(L-n)}{\alpha L - (\alpha+1)n} \ln[\alpha(L-n)/n] \,. \tag{26}$$

Now, after inserting Eqs. (22) and (26) into Eq. (6), we obtain for the expected evolutionary rate

$$K = 2U \frac{\alpha n(L-n)}{\alpha L^2 - (\alpha+1)Ln} \ln[\alpha(L-n)/n] \,. \tag{27}$$

Note that this expression is independent of the population size N. Even though we have derived it under the assumption that N is large, we find that it works very well even for moderate population sizes of 100 or less.

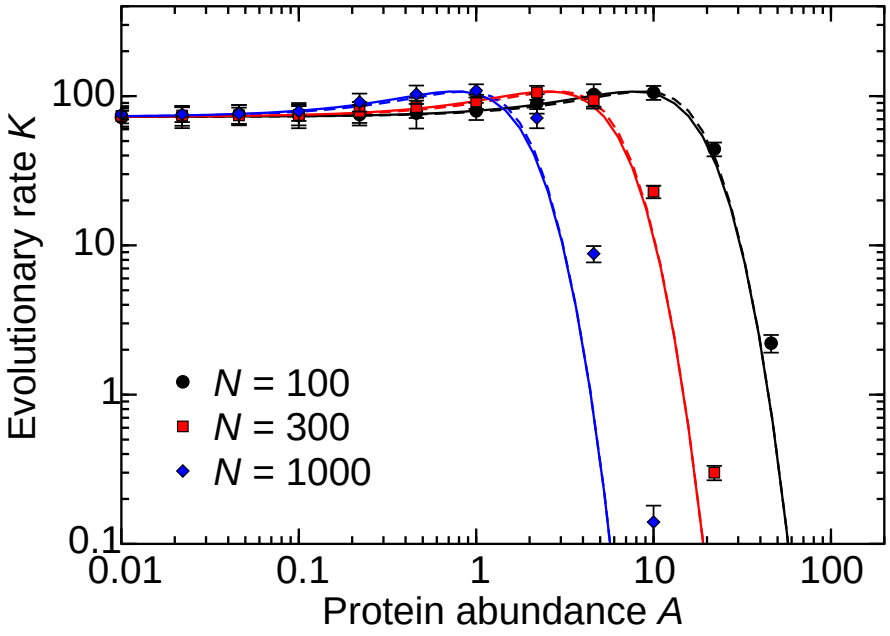

Figure 1: Evolutionary rate K (measured in substitutions per 4×10^5 generations) versus the protein abundance A, for population sizes $N = 100, 300, 1000$ ($\alpha = 0.1$). Data points indicate simulation results, solid lines indicate the prediction using Sella-Hirsh theory, Eq. (6), and dashed lines indicate the prediction using our approximate solution, Eq. (18). Error bars indicate standard errors.

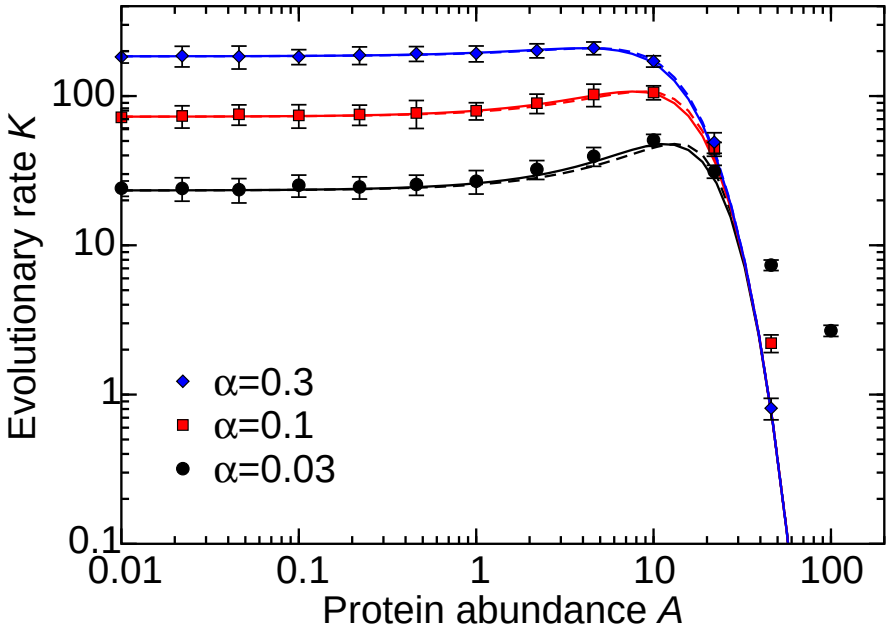

Figure 2: Evolutionary rate K (measured in substitutions per 4×10^5 generations) versus the protein abundance A, for values of $\alpha = 0.03, 0.1, 0.3$ ($N = 100$). Data points indicate simulation results, solid lines indicate the prediction using Sella-Hirsh theory, Eq. (6), and dashed lines indicate the prediction using our approximate solution, Eq. (18). Error bars indicate standard errors. The deviation from the prediction of the two rightmost data points for $\alpha = 0.03$ is caused by insufficient equilibration time.

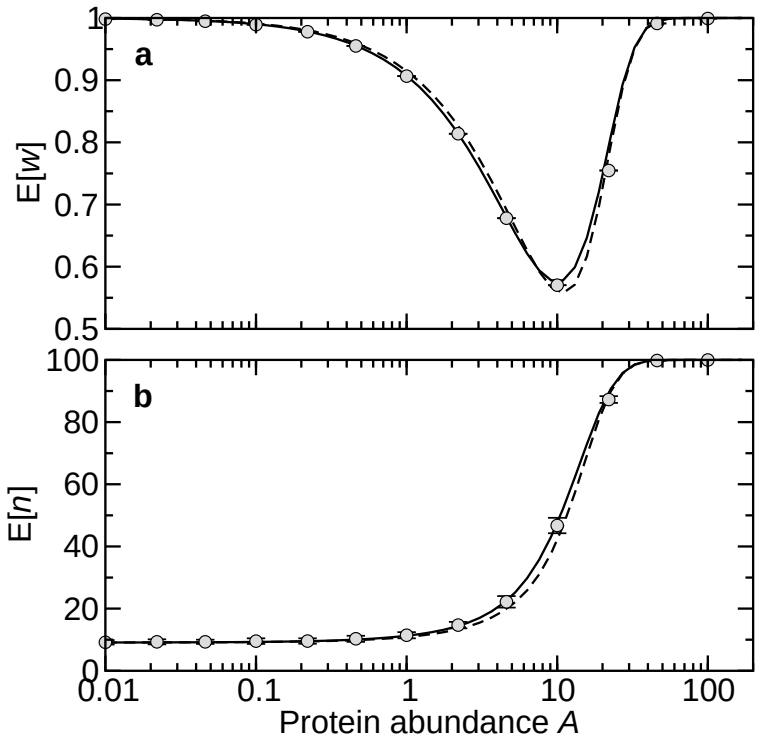

Figure 3: Expected fitness E[w] (a) and expected number of neutral sites E[n] (b) versus the protein abundance A, for $N = 100$ and $\alpha = 0.1$. Data points indicate simulation results, solid lines indicate the prediction using Sella-Hirsh theory, and dashed lines indicate the prediction using our approximate solution. Error bars indicate standard errors.

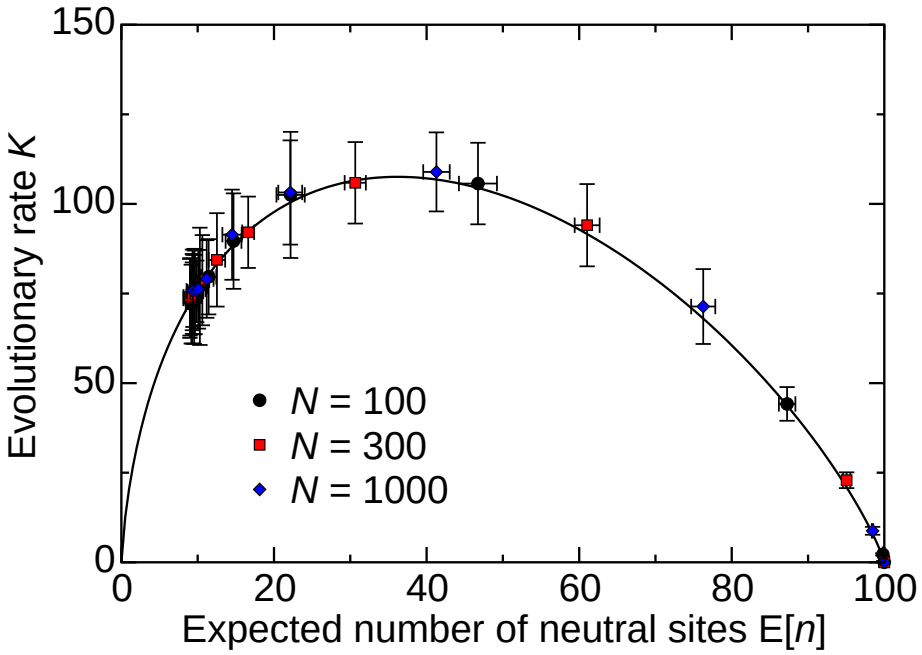

Figure 4: Evolutionary rate versus expected number of neutral sites E[n], for population sizes $N = 100, 300, 1000$ ($\alpha = 0.1$). The solid line stems from Eq. (27) in the Appendix. Error bars indicate standard errors.

α	probability of turning a non-neutral into a neutral site
A	number (abundance) of folded proteins
f	fraction of folded proteins
F	Boltzmann factor
K	number of amino-acid substitutions since most recent common ancestor (evolutionary rate)
l	number of possibly-neutral sites
L	protein length
m	number of never-neutral sites
n	number of neutral sites
N	population size
t	time, in generations
τ	translation error probability per site
U	mutation rate per sequence
w_n	fitness of sequence with n neutral sites

Table 1: Variables used in this work.